Homework 6: Supervised Learning: K nearest neighbours

Marina Sedinkina,Benjamin Roth
Symbolische Programmiersprache

Due: Thursday December 14, 2017, 16:00

In this exercise you will:

- implement k nearest neighbours classifier

Exercise 1: K nearest neighbours [6 points]

Train k nearest neighbours classifier using training set of newsgroups data and classify test documents (test set) into one of the 20 newsgroups.

Download and unpack `20news-bydate.tar.gz` - 20 Newsgroups sorted by date from `http://qwone.com/~jason/20Newsgroups/` into the `data/` folder of your project. The dataset contains train and test folders consisting of several newsgroups folders and their documents. Take a look at the data and the file `hw06_knn/classification.py`. In this exercise you will have to complete some methods to make the classification work.

This homework will be graded using unit tests by running:
`python3 -m unittest -v hw06_knn/test_knn.py`

1. Complete the method `choose_one(self, labels)`. This method should return unique neighbor (label) from the given k nearest neighbors (labels). If there is a unique winner, return it, otherwise, reduce the number of k and search again.
2. Implement methods included in `classify(self,test_files)`:
   a) `calculate_similarities(self,test_doc,train_docs)`: calculate similarities between test document and other train documents; do not forget to label them ([(similarity, label),...])
   b) `order_nearest_to_farthest(self,similarities)`: order the pairs of similarity and label from most similar to less similar

c) `labels_k_closest(self,sorted_similarities)`: find k closest labels

d) `append_pred_label(self,results,k_nearest_labels)`: append winner label to the results

3. Implement the method `get_accuracy(self,gold_labels, predicted_labels)`. This method should return the accuracy: proportion of correctly classified test documents over the whole test set of documents